Users Guide to Designing N-of-1 Trials

Chapter 4

Statistical Design and Analytic Considerations for N-of-1 Trials

Draft for Comment

Chapter 4. Statistical Considerations for N-of-1 Trials

We discuss in this chapter key statistical issues for N-of-1 trials – trials of one patient, treated multiple times with two or more treatments, usually in a randomized order, with the design under the control of the patient and his or her clinician – including special features of their experimental design, data collection strategies and statistical analysis. For simplicity, we will focus on the two-treatment, block pair design in which patients receive each of two treatments in every consecutive pair of periods with treatment assignments occurring separately within each block of two periods, either randomized or in a systematic, balanced design. Extensions are straightforward to other designs such as K treatments (K > 2) assigned in blocks of size K, randomization schemes with different sized blocks (e.g., block sizes equal to a multiple of the number of treatments), or unblocked assignment schemes, requiring no changes in the fundamental principles we outline. The basic design principles include randomization and counterbalancing, replication and blocking, the number of crossovers needed to optimize statistical power and the choice of outcomes of interest to the patient and clinician. Analyses must contend with the scale of the outcomes (continuous, categorical or count data), changes over time independent of treatment, carryover of treatment effects from the one period into the next, (auto)correlation of measurements, premature end of treatment periods and modes of inference (Bayesian or frequentist). All of these complexities exist within an experimental environment that is not nearly as carefully regulated as the usual randomized clinical trials and so require an appreciation of the special difficulties of gathering data in an N-of-1 trial.

5.1 Experimental Design

N-of-1 studies are appealing because they allow the patient and clinician to devise an individualized trial with idiosyncratic treatments and outcomes run in a non-traditional research setting. As a result, N-of-1 designs may vary substantially and may be quite creative. On the other hand, they often involve clinicians unfamiliar with the principles and practice of clinical trials and who may not have access to the resources common in research settings. Because many N-of-1 trials will be carried out in non-research medical office or outpatient clinic environments, it is important to ensure that proper experimental standards are maintained while at the same time allowing designs to remain flexible and easy to implement. One way to ensure such standards is to establish a centralized service responsible for crucial study tasks such as providing properly randomized treatment sequences to the patient-clinician pair when they are designing the trial. We next discuss common clinical crossover trial standards that continue to apply in N-of-1 studies.

Randomization/Counterbalancing After choosing the identity and duration of the treatments to be given, the patient and her clinician must be given a sequence of treatments in such a way that the validity of the experimental process is maintained. The sequence can be either randomized or generated in a systematic counterbalanced design, such as ABBA [1,2]. In the standard two-treatment N-of-1 trial, the assignments are made within blocks of two time periods. With randomization, the first time period in each block is assigned randomly to one of the two treatments, say, A; the second time period is then assigned to the other treatment, say, B. With a counterbalanced design, the assignments alternate between AB and BA, in a systematic manner that is expected to minimize possible confounding with time trend. For example, each two blocks can be assigned as AB (first block) BA (second block), to eliminate possible confounding with a linear time trend.

An important requirement for a good experimental design is to balance treatment assignments, especially for potential confounding factors, so that the treatments are compared fairly. Making assignments in blocks of size two ensures that each patient receives each treatment with the same frequency at a comparable set of times, to avoid poorly balanced designs like AABA and AABB.

Randomization and counterbalancing both attempt to balance treatments within and across blocks. Randomization achieves balance on expectation, when averaged across a large number of blocks, and/or a large number of N-of-1 trials. For each individual N-of-1 trial, exact balance might not be achieved. For example, if patient outcomes might be deteriorating gradually over time inducing a time trend, the ABAB design would not be well-balanced as B is always delivered after A. The design itself may induce inferior outcomes for B due to the time trend when the two treatments are actually equivalent. For a four-period trial randomized in blocks of size two, there is a 50% chance for randomization to yield such an unbalanced design, either ABAB or BABA (and 50% chance to yield a design that is well-balanced against the linear time trend, either ABBA or BAAB). Counterbalancing, on the other hand, can be more effective at achieving exact or nearly exact balance for the potential confounding factor(s) designed explicitly to be balanced, e.g., the ABBA design achieves exact balance for linear time trend.

While randomization can be less effective than counterbalancing in balancing for known confounding factor(s), randomization has an important advantage in its ability to balance (on average) all potential confounding factors, both known and unknown. Counterbalancing, on the other hand, can behave poorly if the explicit scheme chosen leads to imbalance with respect to an unknown confounding factor.

In addition to reducing but perhaps not completely eliminating the risk of bias induced by time trends, blocked assignment also provides two other important benefits. It minimizes the consequences of early termination of the trial that might otherwise lead to an unbalanced number of observations in the two treatment arms. Within-block assignment also reduces the chances that unknown confounders may bias the estimate of within-patient variation that would invalidate appropriate statistical inference.

To summarize, we recommend that a blocked scheme for treatment assignment be used for N-of-1 trials. We also recommend that users make a careful choice between randomization and counterbalancing. If there is good information on the most important potential confounding factor (such as the linear time trend), counterbalancing can be more effective. Otherwise, randomization would be a more robust choice. The end of the section on Blinding has some further discussion.

Blinding To the extent possible, patients and clinicians should remain blinded to the treatment assigned, particularly when patient-reported or other subjectively ascertained outcomes are used. While blinding is desirable in all clinical trials, it may be particularly important with N-of-1 trials because of the individualized crossover nature of the study. Patients may (and probably will) naturally try to guess which treatment they received in each period. Because they are so invested in the research and so desirous of a positive outcome, it is natural that their reported outcome measures may be subtly and unconsciously (or not so subtly and unconsciously) affected by knowledge of the treatment received, e.g., in favor of the direction that confirms any pre-existing expectations they might have (the effect of expectancy). On the other hand, patients' self-interest might also drive them to report as objectively as possible, particularly if they enter the trial without any preconceived preferences, because they themselves will bear the consequences of a bad treatment decision made because of biased outcome reports. Potential bias might also ensue from the motivation for the trial if, for example, patients were compelled to enter an N-of-1 trial to prove that a more expensive treatment was really indicated and should be reimbursed.

In the absence of pure blinding, other features related to the treatment administration might influence outcomes, but in such a way that they should actually be incorporated into the treatment decision if it is reasonable to expect the same effect will sustain beyond the end of the N-of-1 trial. For example, if the patient prefers the one pill to the other because of its color or texture during the trial and this effect can be sustained in the future, it is a real effect for this patient and should be part of the treatment decision. In a parallel group trial where the intent is to generalize beyond the patients in the trial, such a preference should be considered a bias, because future patients to be treated according to the findings from the trial might not have the same preference for the same type of pill. In addition to the potential effect on reported outcomes, knowledge of treatment identity may lead some to end a treatment period early if the measured outcomes support the treatment expectation. Even if the treatment assignment is blinded, superior results in one or more periods may induce patients to ask to unblind the trial to see whether their hunches are correct. Such unblinding will stop the trial and may result in an inconclusive final result.

For blinded N-of-1 trials with treatments assigned in small blocks such as blocks of size two, there is sometimes a concern that some users (patients and/or clinicians) might learn during the course of the trial that the second treatment in the block is predetermined by the first, therefore the outcome for the second treatment might be affected by the effect of expectancy. When this is an important concern, one could use a block size that is a multiple of the number of treatments, or randomize the block sizes in different multiples of the number of treatments. This strategy minimizes the chance for the user to figure out the treatment in any given period. On the other hand, this strategy may also increase the risk of bias if time trends are present or dropout occurs.

Replication Because only one patient is involved in an N-of-1 trial, the number of measurements taken on each individual determines the sample size of the study. This sample size comprises two components: the number of treatment periods and the number of measurements taken within each period. For instance, a pain outcome measured daily over six 14-day treatment periods will have 84 observed data points. These repeated measurements enable estimation of between and within-period variance, crucial for proper statistical modeling. Larger sample sizes can be achieved by increasing the number of treatment periods, increasing the length of each period, or increasing the frequency of measurements within each period. These alternative strategies have different analytic implications because they affect different components of the study variance. It is important to carefully choose both the number of crossover periods and the number of measurements taken per period) to enhance the efficiency of the study design. More data will improve the precision of the treatment effect estimate, but the optimal allocation to more treatment periods or more measurements per period depends upon statistical considerations such as the expected size of each variance component and its influence on the precision of the effect of interest and the minimum effect size of interest, as well as on practical considerations related to feasibility and type of measurement. Such considerations include patients' inability to record data more than once a day, lack of measure validity on different time scales, increased likelihood of dropout with longer trials and the tendency for patients to become less careful to follow treatment protocols over time. Outcomes with substantial measurement variation such as quality of life measures will need to be collected more frequently in order to precisely estimate the variance.

Washout Carryover, the tendency for treatment effects to linger beyond the crossover, when one treatment is stopped and the next one started, threatens the validity of the comparison between treatments in crossover studies, including N-of-1 trials. While statistical models may attempt to accommodate carryover, they rely on assumptions about the nature of the carryover that may be difficult to test or even control. In the extreme, carryover may extend throughout all or most of the next treatment period, contaminating many of the outcome measurements.

Inserting a washout period in which no treatment is given between consecutive treatment periods is the common method to reduce or even eliminate the effect of carryover by design. The goal of a washout period is to provide time for each patient to return to his baseline disease state, unaffected by preceding treatment. Deciding whether to include a washout period depends on both clinical judgment about the durability of the treatment effect as well as practical considerations related to satisfaction among end users (patient and clinician) with the study design.

An important clinical consideration for the washout is to avoid adverse interaction between the treatment conditions. This is mainly an issue for active control studies, with an active treatment (the standard treatment) used as the control condition to evaluate the comparative effectiveness of an alternative treatment. If the two active treatments being compared are not compatible with each other, it would be necessary to impose a washout period to eliminate the first agent before starting the second agent.

When adverse interaction can be ruled out, the inclusion of a washout period can be problematic for active control studies, both in terms of satisfaction for the end users (patient and clinician), and in terms of clinical ethics. The washout period introduces a third treatment condition, the absence of either active treatment. It is conceivable that the patient might be managing the disease condition adequately with her current treatment, and undertakes the N-of-1 trial to test the possibility that the alternative treatment might be better. It is undesirable, and perhaps even unethical, for the patient to be forced into a period of no treatment that is likely to be inferior to the current treatment. The use of washout in such studies might reduce the willingness for patients to undertake the N-of-1 trial, and increase the chance for early termination of the trial. The ethical dilemma here is that, when adverse interaction can be ruled out, there is no obvious clinical rationale to withhold both active treatments from the patient during the washout period, other than to make a short term sacrifice in exchange for a better chance to improve the therapeutic precision at the end of the trial.

Conversely, not using a washout might compromise the validity of the estimated treatment effect and lead to biased estimates for treatment effects. Therefore users need to determine whether the likelihood for a substantial bias warrants the drawbacks of the washout.

In some cases, the effect of the washout can be accomplished analytically without including any period during which treatments are withheld. More specifically, any effect of carryover can be dealt with analytically by eliminating, discarding or downweighting observations taken at the beginning of a new treatment period. It is also possible to introduce a time-to-respond statistical model to include all observations while allowing a carryover effect to be included in the model as a transient function that drifts towards zero gradually over time as a smoother method to reduce the influence for potentially contaminated observations early in the period. This approach can help to maintain the integrity of the trial by reducing the chance that the patient will drop out and that observations will be contaminated by carryover.

While carryover affects how the treatment effects of the previous treatment might linger on after the completion of the previous treatment period, another important transition issue is the onset of the new treatment. Some treatments, such as selective serotonin reuptake inhibitors (SSRIs) may take awhile to reach full effectiveness. Slow onset provides another reason to reduce the influence for potentially contaminated data at the beginning of a period; it introduces a natural washout, particularly if the time for one drug to wear off is no greater than the time for the next drug to take effect.

It should be noted that a washout period does not mitigate the problem of slow onset directly. On the contrary, a washout period further extends the transition between the two treatments, because the onset for the new treatment does not begin until the end of the washout period. As an example, assume that treatment A takes three days to washout, and treatment B takes two days to reach its full effectiveness. If a washout period of three days is used after a period of treatment A, then treatment B begins on day four and reaches its full effectiveness on day six. Therefore, a total of five days are lost to the transition between the two treatments. On the other hand, if a washout period is not used (under the assumption that there is no adverse interaction between the two treatments), the transition is three days only: by day three, treatment B has reached its full effectiveness; by day four, the carryover effect for treatment A has disappeared. Therefore only three, instead of five, days of treatment do not reflect full treatment effects.

If a washout period is included in the study design, choice of its length needs to be made carefully, taking into consideration treatment interactions, medical ethics, drug half-lives and onset efficacy. Longer washout periods decrease the likelihood of carryover, but increase the length of the study and time spent off treatment, and also delay the onset of the full effectiveness of the next treatment. Making washout periods too short contaminates treatment effects and carryover effects, and might result in biased estimates for treatment effects. In summary, one needs to define treatment periods sufficiently long to manifest an effect, but short enough to allow enough crossovers within a reasonable total duration for the study.

Adaptation While a fixed trial design is the norm, adaptive trial designs offer the chance to modify the design of a trial in progress in order to make it more efficient or to fix problems that may have arisen [3]. Some adaptations occur naturally, as when a patient and clinician decide to stop a trial because one treatment appears to be more effective or end a treatment period early because of an adverse event. It is important in such circumstances that blinding be maintained if it is already part of the study design. For instance, it would not be proper to unblind a treatment period in order to stop one treatment, but not the other. Other adaptations could include extending the length of the trial to more treatment periods if treatment differences appear to be small or instigating play the winner designs [4,5] where the treatment that appears to be more effective is given more frequently. Such designs are generally easier to implement when the data are analyzed using Bayesian methods without tests of hypothesis whose properties depend on prespecified design plans. If frequentist inference (i.e., p-values) is used, one needs to use sequential design with explicit stopping rules so as to protect the overall type I error rate. In some cases, decisions to adapt a design may arise from experience with similar patients.

Multiple Outcomes The personalized nature of N-of-1 trials and their focus on making a treatment decision for an individual patient require outcomes to be carefully chosen so as to reflect the measures of most importance to the patient’s well-being. Often, more than one outcome is of interest to the patient – perhaps reducing pain and sleeping better – and so the effect of treatment on both needs to be considered in making the choice of treatment at the end of the trial. This contrasts with most clinical trials, which often focus on one particular average treatment effect in the population. Thus, although almost all clinical trials collect data on at least several, if not many, outcomes of interest, they typically focus on a primary outcome and so use statistical methods for a single outcome variable.

A common technique when multiple outcomes are of interest is to form a composite variable such as MACE in cardiovascular trials, which counts the number of Major Adverse Cardiac Events (e.g., acute myocardial infarction, ischemic stroke, coronary arterial occlusion and death), and then analyze it by univariate methods. Composite outcomes are not as popular in N-of-1 studies because they do not allow the patient or clinician to see the effect on each distinct outcome separately. Often, the outcomes differ so fundamentally that forming a composite becomes difficult. Returning to a previous example, how might one combine a pain scale and the number of nights of good sleep over a fortnight? One could express both as a percentage of relief compared to a baseline level and then average the two percentages, but this would assume that each outcome was of equal importance and that both outcome scales were linear. Alternatively, one could choose one as primary and the other as secondary, but if the patient were concerned with both then this is unlikely to work well. Another approach would be to form a weighted composite scale, with weights accommodating patient and clinician preferences or utilities.

To reflect the patient's true decision-making state, one might instead analyze each outcome separately and report a measure of the treatment's effectiveness for each, letting the patient and clinician weight them on their own. One could argue, however, that explicitly specifying the weights up front is more scientific and transparent than having the patient and clinician implicitly weighting separate outcomes in trying to make a treatment decision. In the end, this is a decision problem and it is worth exploring methods of decision analysis to improve decision making for N-of-1 trials. Both approaches may be useful.

Because the focus is on the immediate decision of which treatment to take, it is not important to protect against a false positive decision as in the standard test of hypothesis commonly employed in clinical trials. One is not choosing to report a statistically significant finding for one outcome among many, so multiple testing is not an issue. Instead, one provides the decision maker with all the information required to make the decision in a format that facilitates decision-making.

Multiple Subjects Designs Several publications have described an N-of-1 service in which many patients are offered the opportunity of carrying out studies. Such services offer several advantages: economies of scale in research infrastructure, clinicians experienced in N-of-1 trials and the chance to use information gained from other patients. Multiple N-of-1 trials may be combined in a common statistical model to both estimate the average treatment effect as well as improve individual treatment effect estimates by borrowing strength from the information provided by other similar patients. As more patients accrue, not only does the precision with which the next patient can be evaluated improve, but also the estimates for previous patients that might have even finished their studies may change as a result of information gathered on later patients. Multiple subject designs increase the complexity of sample size choices because they permit manipulation of the number of subjects as well as the number of measurements on each. Balancing these two numbers requires knowledge of the relevant within and between-patient variances [6]. Ethical considerations may also arise from multiple N-of-1 trials if one treatment appears to be working better and clinicians become reluctant to continue randomizing patients due to lack of equipoise.

5.2 Data Collection

The lack of research infrastructure for the single clinician running an N-of-1 trial may have a serious detrimental effect on data collection. Typically, research studies initiate elaborate procedures to ensure that data are collected in a timely, efficient, accurate fashion. Forms are tested and standardized; research assistants are hired and trained to help collect data from patients either at patient visits or remotely via mail, telephone or internet connections; data are checked and rechecked by trial personnel and external monitors; and missing items are followed up. Many of these options may not be available to the typical clinician running a trial outside of an established N-of-1 service. Conversely, patients in N-of-1 trials are usually extremely motivated because the trial is being done for them and by them, so they may be more committed to data collection and therefore less likely to miss visits and fail to complete forms accurately. Missing items can be particularly costly in an N-of-1 study because of the small number of observations.

Clinicians undertaking N-of-1 trials must be aware that each trial is unique, with its own protocol and its own set of outcomes. This multiplicity of designs can complicate data collection, even if a service is available. Multiple data collection forms may be needed and personalized user interfaces may be valuable ways to collect data. Reminders are important to provide and interim feedback can maintain the patient's enthusiasm.

5.3 Statistical Models and Analysis

The unique design features of N-of-1 trials, including a multiple period crossover design, multiple patient-selected outcomes and focus on individual treatment effects, motivate statistical models for these trials. Data resemble a time series in that they are autocorrelated measurements on a single experimental unit. Unlike classic time series, however, the measurements are structured by the randomized design and so statistical models also have features like those for longitudinal data with a time-varying covariate (the treatment condition). The main goal is to compare the observations taken under the two treatment conditions, adjusting for any carryover effects, while accommodating the randomized block structure.

Constructing such models is difficult, especially when few measurements are taken. A review of the N-of-1 literature in medicine in fact found that many studies have used no formal statistical model at all to compare treatments, opting instead for eyeball tests based on a graph of the data or simple nonparametric tests such as the proportion of paired treatment periods in which A outperformed B [7]. When the data are simple and treatment differences are clear, such simple methods work well. Graphs are always informative and plots of the measurements provide good ways to understand the data. But when the number of measurements gets large or when differences are small, graphs will not be sufficient to properly distinguish the treatment effects.

The basic data from an N-of-1 design consist of measurements taken over time while on different treatments. The fundamentals of the statistical analysis can be most easily understood by focusing on the two treatment design in which treatments are randomized in blocks of size two, each treatment appearing once in each block. Each treatment period consists of one or more measurement times.

Nonparametric Tests The earliest N-of-1 trials in medicine used a simple type of nonparametric test called the sign test. First, one calculates the difference between treatment A and treatment B. If the difference is positive (A is better than B), one counts this as a success. A negative difference counts as a failure. (The choice of which difference is defined to be a success is, of course, arbitrary). The number of successes, i.e. the number of blocks in which A outperforms B, is now compared to the number expected if the treatments were the same which is N/2 where N is the number of blocks. Since the number of successes is assumed to follow a binomial distribution, one calculates the probability of the observed result under the null hypothesis that the true success probability is ½. For example, if there were three blocked comparisons and in each A was better than B, the probability would be ½*½*½ = 1/8. This is then a (one-sided) p-value for testing whether A was better than B. This procedure ignores the actual size of the differences and thus throws away potentially important information. Instead, one might use the Wilcoxon signed-rank test on the ranked differences.

While these simple nonparametric tests are easy to use, they ignore important features of the time series data, particularly their autocorrelation, time trends and repeated measurements within periods. As a consequence, it is usually worth constructing a proper statistical model that incorporates these features along with an estimation of treatment effect.

Models for Continuous Outcomes A variety of different models can be constructed when the outcomes are continuous variables depending on whether they are considered random measurements within each treatment period or vary systematically with time.

First, consider a model in which time may be indexed within treatment periods inside blocks. Notationally, let y_{ijkl} represent the outcome measured at time i within treatment period j within block k while on treatment l:

Model 1: $y_{ijkl} = \alpha + \beta_l + \gamma_k + \delta_{j(k)} + \varepsilon_{i(j(k))}$.

Model 1 assumes a fixed treatment effect β_l, random block effects $\gamma_k \sim N\left(0, \sigma_\gamma^2\right)$, random period within block effects $\delta_{j(k)} \sim N\left(0, \sigma_\delta^2\right)$ and random within-period errors $\varepsilon_{i(j(k))} \sim N\left(0, \sigma^2\right)$, where the notation $N\left(\mu, \sigma^2\right)$ indicates a normal distribution with mean μ and variance σ^2. The constant term is used to avoid oversaturation of model terms. Usually, one block is chosen as the reference (e.g., set $\gamma_1 = 0$) and period within block effects may be expressed so that the difference between the first and second period is assumed the same in each block. This model assumes no time trend and no carryover. The model may be simplified if observations within one treatment period or block are uncorrelated with those in another. In that case, the model becomes a simple two mean model with random errors

Model 2: $y_{ijkl} = \gamma_l + \varepsilon_{i(j(k))}$.

Source: http://www.effectivehealthcare.ahrq.gov/index.cfm
Date Posted: May 20, 2013

A common scenario for this model would occur if each treatment period had only one observation, perhaps at the end to minimize the possibility of carryover.

Modeling Effects Depending on Time Another class of models pertains to occasions when outcomes vary systematically with time. Causes for such variation include time trends that might describe a disease course or calendar effects that arise from seasonal variation in severity, for instance in asthma patients whose health is affected by hay fever. Measuring such time effects requires that the study duration and measurement frequency be sufficient to differentiate the trends from noise. It is easiest to then express the model in terms of the measurement y_t taken at time t. If the trend is linear, we have

Model 3: $$y_t = \alpha + \beta t + \gamma X_t + \varepsilon_t ,$$

in which β is the slope of the time trend, X_t is an indicator for the treatment received at time t, γ is the treatment effect and ε_t are the residual errors possibly correlated over time. Other calendar effects can be introduced by modifying the time variable. For instance a seasonal effect could be introduced by adding a dummy variable Z_t taking the value one during the season and zero outside it. When each period has a single measurement, the time variable can be replaced by an indicator variable for period. If the effect of treatment is expected to vary with time (e.g., because of higher efficacy during periods of greater disease severity), one can include a time by treatment interaction effect into the model.

Autocorrelation Measurements in a time series typically are not independent, exhibiting some form of autocorrelation that represents the relationship between one measurement and the next in the series. Such autocorrelation arises from time trends or treatment carryover that causes individuals to tend to respond more similarly at times that are closer to each other. Model 3 presents one method of detrending the time series by fitting a model linear in time. Such detrending often removes substantial amounts of observed autocorrelation, but some may remain as a consequence of features like carryover or delayed uptake. Carryover occurs when the effect of a stopped treatment continues into the next period after a new treatment is introduced. It will cause the response to be greater than it should be, effectually because two treatments are acting instead of one. Delayed uptake applies if the full effect of a treatment is not felt at the start of the measurement of the outcome. It will work in the opposite direction, depressing the response initially. The effect of each, however, is to induce correlation between consecutive outcome measurements.

Models that adjust for autocorrelation take two main forms. The first, often called an autoregressive or serial correlation model, expresses the residual error at a given time as a function of the error at one or more previous times, i.e., $\varepsilon_t = \delta\varepsilon_{t-1} + u_t$. In this model, δ is the correlation between consecutive errors ε_t and ε_{t-1}. Additional lagged errors of the form ε_{t-k} can be added to the model to represent more complex autocorrelation. The second form, called by some a dynamic model [8] places the autocorrelation on the outcomes themselves so that the response at time *t* is a function of the response at time *t-1* (and perhaps earlier times). A dynamic form for a model with one fixed treatment effect, for instance, would be $y_t = \delta y_{t-1} + \gamma X_t + \varepsilon_t$. The dynamic model induces a dependence of the current outcome on previous values of the predictors in the model. One can also explicitly introduce this dependence by introducing lagged predictors. It is important to recognize the different interpretation of predictors in a dynamic model resulting from the need to condition on the previous outcome, i.e. g is the treatment effect conditioning on y_{t-1}.

Carryover Carryover is a special type of autocorrelation common to crossover trials. It occurs when the time between treatment periods is insufficient for the effect of the previous treatment to end before the next treatment is started. This is common with pharmacological treatments when the drug continues to metabolize in the body for a period of time after the patient stops taking it. If not controlled for, carryover may lead to bias in the estimated treatment effects, with a tendency to magnify observed treatment effects during transitions from a less effective (but still effective) treatment to a more effective treatment, and conversely to shrink effects during transitions from a more effective to a less effective treatment.

Both design and analytic approaches can address carryover. Designing washout periods long enough for the prior treatment's effect to disappear by the beginning of the next treatment period eliminates any potential correlation across periods. An analytic approach downweights, disregards or does not collect outcomes at the beginning of a treatment period, thus creating an analytic washout period [9]. This analytic approach is also helpful when treatments take time to reach their full effect and it is desired to account for the reduced effect at the beginning of the period.

Zucker [10] used an extreme version of this approach in a series of N-of-1 trials for patients with fibromyalgia tested on amitryptoline or amitryptoline plus fluoxitene. Treatment periods were six weeks long and the primary outcome was the score on the patient-reported Fibromyalgia Impact Questionnaire. Only the report from the end of each treatment period was analyzed. While this almost certainly eliminated carryover, and in fact autocorrelation, it did have the drawback of giving only one measurement per treatment period. In some studies, however, these choices may be unavailable if each treatment period is short or treatment half-life is very long.

Various approaches to estimating carryover have been proposed. As Senn [11]points out, all rely on restrictive modeling assumptions and are inferior to designing a proper washout (which also may rely on assumptions about pharmacologic or similar properties of the treatments). The discussion above points to autocorrelation models as one method to handle carryover, although they assume correlations over time unrelated to when treatment is changed or introduced. One could, in principle, design an autocorrelation structure that varied with time since introduction of treatment. But this would need to assume knowledge of the nature of the carryover that might not be well supported.

A simple check for carryover when the analyst has a sufficient number of observations taken over time within each treatment period is to compare results using all measurements and after having discarded those at the beginning of the period that might be affected by carryover. The model with more measurements should return more precise estimates but at the risk of some bias from the carryover. If the estimates are similar, carryover is not likely to be an issue.

Another from of carryover that one might be able to examine is the effect of treatment sequence when the response is different depending on the order of the treatments given. Treatment A may have a bigger effect if given after treatment B. This might manifest itself through responses that are higher for treatment A when it follows B than when it follows another period of A. One can examine a sequence effect by adding a variable that codes for sequence, e.g. a dummy variable that equals one in periods where A follows B and zero otherwise. Of course, if treatment effects are wearing off, it would not be appropriate to code every measurement in the A period with the sequence effect.

Discrete Outcomes In each of the models presented, we have assumed a continuous outcome with normally distributed measurement error. Many outcomes that might be used in N-of-1 trials, however, may use categorical scales, event counts or binary indicators of health status. For example, Guyatt [12]and Larson [13] both used Likert scales with ratings from 1-7 to measure patients outcomes. Models for such outcomes require different formulations that do not rely on the assumption of normality.

Generally, one needs to formulate such models as generalized linear models [14]. Binary outcomes use logistic regression; count outcomes use Poisson regression; and categorical outcomes use categorical logistic regression. The generalized linear model has the same form as the linear model on the right hand sides of the models above, but expresses the left-hand side in terms of a (link) function of the mean of the probability distribution for the outcomes. For example, with a binary outcome, events occur according to Bernoulli distribution and the mean of that distribution is the probability of an event. The link function used in logistic regression is the logit function (logit (p) = $\log_e(p/(1-p))$). In Poisson regression, the link function is log. For categorical regression, various link functions can be used depending on how one wants to model the data. A common link function for an ordered outcome such as a preference scale is the cumulative logit [15].

Although the generalized linear models use different estimation algorithms and take different functional forms, model construction does not differ conceptually in any fundamental way from the normal linear models, so we will say no more about them here, but refer the interested reader to the many textbooks that treat them[14,15] .

Estimation The simplest approach to estimating the treatment effect uses the model that ignores any potential effects of time, autocorrelation or carryover and simply compares the average response when the patient is on each treatment. If the design is blocked, one can take the difference between outcomes within each block and then simply average the differences computing the appropriate standard error. This corresponds to a paired t-test. If no blocking is used, the analysis is an unpaired t-test[11].

In general, one can use likelihood or Bayesian methods that incorporate the necessary correlation structures and interaction terms to fit the models. Likelihood-based methods typically rely on large samples to validate their assumptions of normal distributions of the resulting model estimates. Because the amount of data collected on any single outcome in an N-of-1 study is small, such assumptions may not be appropriate.

Bayesian inference combines the likelihood with prior information to form a posterior distribution of the likelihood that a model parameter takes a given value. The prior information is expressed through a probability distribution describing our degree of belief about model parameters before observing the data. Bayesian inference is natural for clinicians making decisions such as a differential diagnosis because it expresses the way that they combine new information (such as a diagnostic test result) to update their previous beliefs [16]. The use of prior information also permits the analysis to incorporate patient preferences and beliefs.

Specification of a complete prior distribution for all model parameters can be difficult, particularly for those like correlations or variance components about which not much may be known. One common simplification assumes that very little is known about some or all of the parameters and uses prior distributions that do not favor any values over others. Probabilistically, this corresponds to a uniform (flat) distribution. Such priors are called noninformative. Conversely, knowledge of certain parameters such as the expected treatment effect may be available and so informative priors may be chosen. For example, for a pain scale outcome the average amount of pain reduction that one can expect over a two-week course of therapy may be approximately known in the population or one may be able to bound the maximum amount. It is also possible to construct an approximate prior distribution by eliciting some of its features, such as means or percentiles [17].

The posterior distribution, formed by calculating the conditional probability distribution of each parameter given the observed data and the specified prior distribution, is essentially a weighted average of the observed treatment effect mean and the hypothesized prior mean. The weights are supplied by the relative information about the two expressed through the precision with which each is known. One can use the posterior distribution to make statements about the probability that the parameters take on different values. For instance, one might conclude that the chance that treatment A reduces pain more than treatment B as measured on a specific pain scale is 75%; or, one might say that there is 50% chance that the reduction is at least 10 points on the scale. Statements like this can be made for each outcome, allowing the patient and clinician to weigh them and determine which treatment is working better. Bayesian inference leads to statements about the probability of different hypothesis given the data observed; non-Bayesian, or frequentist, inference leads to statements about the probability of the data given the null hypothesis.

Local Knowledge and statistical methods The personalized nature of N-of-1 trials indicates that the primary use for the knowledge produced in each individual trial is to inform clinical decision-making for the specific patient, i.e., the knowledge produced is used locally or internally within the patient-clinician team that produced this knowledge. This paradigm is crucially different from the situation in the standard parallel group RCTs, in which the primary use of the knowledge produced in an RCT is to inform clinical decision-making for future patients, rather then for the patients participating in the RCT. In fact, for double blinded RCTs, the patients and their clinicians do not know the treatment the patient actually received until the RCT is unblinded. Given this fundamental difference between the two paradigms, the appropriate statistical method also differs. While significance testing is the usual statistical method for the standard parallel group RCTs, the same method might be less pertinent for N-of-1 trials. Instead, one provides the decision maker with all the information required to make the decision in a format that facilitates decision-making.

Presentation of Results

In order to make a correct decision, it is important that the patient and clinician not only have the right information, but that it be presented to them in a format easy to understand. The results of a trial are complex. Data are collected on multiple outcomes at multiple times under different treatment conditions. Many of the models we have discussed describe complicated phenomena like autocorrelation that may confound facile interpretation of the data. Good graphics can help explain the data and the results to both parties.

The simplest graph that should always accompany results plots each outcome over time separately in the treatment and control groups. A variety of different approaches are possible. One could overlay or stack two line plots, matching by block pairs. This reveals within-block differences as well as time trends and potential autocorrelation. One could add the sequence order by separately coloring within each block the first sequence in one color and the second in another as in Figure 1. Displays of the raw data, like Figure 1, provide important information on the relationship of outcomes to treatment. They may also be shown in a blinded fashion (without identification of treatment group) to the patient during her trial as a form of patient feedback to motivate adherence.

Determining treatment differences directly from such figures may, however, be camouflaged by other features of the data like autocorrelation and time trends. Figure 1 shows simulated data that appear to show that treatment B (dotted line) typically produces higher outcomes than treatment A (solid line). Responses appear to be increasing with time on treatment A, but not B, suggesting a potential treatment by block interaction. Because only one measurement is recorded on each treatment period, we cannot distinguish time effects from effects by block. The overall effect of the picture is that B may be better than A, but that this efficacy wears off over time. In fact, the data are simulated with a treatment difference and with a trend over time, but no treatment by block interaction, which occurs by chance. The right answer is that B is better than A and that all patient responses are increasing with time. Therefore, the plot is somewhat misleading and may lead to the wrong decision. As a general rule, unless treatment effects are large or specific, plots will provide necessary, but not sufficient information to make appropriate decisions. It is therefore important to supplement the graphs with appropriate statistical analysis and present the information in the clearest way possible.

One should use the statistic provided by the modeling process that relates directly to the measured treatment difference. In the Bayesian framework, this is the posterior probability; in the non-Bayesian, or frequentist, framework, this is typically a p-value. We recommend the Bayesian approach because it provides more value to the patient. The p-value describes the likelihood of the data under a specific null hypothesis. For example, a p-value of 0.05 for a test of the null hypothesis of no difference in treatments means that if the two treatments had the same effect, one would have observed the difference found 1 in 20 times under repeated sampling. Putting aside the irrelevancy of the repeated sampling assumption since the experiment will not be repeated, one is left with the observation that it is unlikely that the treatments have the same effect. But one does not know the likelihood of any other effect.

Contrast this with the Bayesian interpretation, which gives the full posterior probability distribution of the treatment effect under the model chosen. From this posterior distribution, one can make probabilistic statements about the likelihood of any size of treatment effect, for example the likelihood that the treatment effect is at least 10, or between 5 and 15. In essence, this approach focuses on estimation of the magnitude of the effect, rather than on hypothesis testing.

This focus can be particularly informative when multiple outcomes are of interest to the patient and one wants to balance different objectives. As an alternative to the use of the composite scale discussed previously, one can formulate a joint posterior distribution to make probabilistic statements about the joint probabilities attached to combinations of the outcomes if one were prepared to make some assumptions about their relationships. As an example, assume that the user (patient and her clinician) specified a performance target for the new treatment, A, to improve pain by at least 10 percent and increase sleep by at least one hour per night, compared to the current treatment, B. In the simple (and perhaps unrealistic case) that the outcomes are independent, the probability for the joint outcome is the product of the probabilities of each separate outcome. So, if the probability that A improved pain by 10 percent was 0.3 and the probability that A increased sleep by one hour was 0.2, then the probability that both would happen would be 0.06.

Such probabilities can be expressed by a distribution function of the likelihood of each gain or by a cumulative distribution. As an example, assume that the posterior distribution of treatment benefit on A compared to B for outcome A expressed as a difference in percent change from baseline was normally distributed with mean 10 percent and standard deviation 5 percent. Therefore, there is roughly a 97.5 percent probability that A has bigger benefit than B since 0 change is about 2 standard deviations below the mean. Likewise, assume the benefit for the second outcome is smaller but more uncertain, normally distributed with mean 5 and standard deviation 10. Figure 2 (top row) plots the resulting posterior probability distributions of treatment effect for each outcome together. One might also be interested in their cumulative distributions, or more likely, the probability of observing an improvement at least as big as a certain size. These graphs appear in the middle row of the Figure. Using the dotted lines on the graph, we can see that the probability of at least a 10 percent improvement is slightly higher with outcome 1 than with outcome 2 since its mean is higher, but that the situation reverses for the probability of at least a 20 percent improvement because of the greater uncertainty associated with outcome 2. The bottom row of the figure gives the probability that both outcomes are improved by a given amount. This probability is smaller than for either outcome alone and for this example is roughly the product of the two individual probabilities because the two outcomes were simulated independently. In practice, these joint probabilities may be quite similar to or quite different from their components depending upon the correlation between the outcomes.

While plots like those in Figure 2 display the entire distribution of effect sizes together with our uncertainty in estimating them, some may prefer a simpler display with less total information, but perhaps in an easier to understand format. The distributions in the top row of the figure may be collapsed into a median and a central interval displaying the values most likely to occur with a given amount of probability, often 95 percent. One may also choose one or more amounts of improvement for which to display probabilities. Figure 3 displays the median and 95% central interval (from 2.5 to 97.5 percentile) for the treatment effect for each outcome. The associated probabilities associated with improvement of at least 0, 5, 10, 15 and

20 percent for each outcome and both outcomes together can be displayed as in Table 1. The users should be able to specify the exact outcome levels for which they want probabilities computed. These may correspond, for instance, to clinically relevant values as determined by the patient and her clinician in collaboration.

Some users may prefer to consider results as odds, rather than probabilities. Others may prefer different metrics rather than treatment effects. A flexible environment in which the user can request results in different ways that are most comfortable and personally informative is a desired feature of any N-of-1 analytic module.

Combining N-of-1 Studies

Although N-of-1 studies are designed for single patients working with a single clinician to make a single treatment decision, many N-of-1 studies may be similar enough to inform others. Furthermore, the small number of crossovers used in many N-of-1 studies may increase the need to combine the index patient's own data with data obtained from other patients who participated in similar N-of-1 trials to increase the statistical precision available for making decisions about individual patients.

Such similarity may arise from the same clinician testing the same treatments with different patients having the same condition; similar patients testing the same treatments with different clinicians; clinicians within the same clinic practicing in similar ways; examining a common set of treatments in different combinations. In each case, we may think of the set of N-of-1 studies as forming a meta-analysis and attempt to combine them using techniques from meta-analysis such as multi-level random effects models, regression and networks. As an added bonus, combining the results can help estimate the average treatment effect in the population as well as the individual treatment effects for single patients. We give a brief introduction here, but refer the interested reader to related treatments in Zucker [18] and Duan, Kravitz and Schmid[19].

To extend the previous models to multiple patients with N-of-1 studies consider

Model 1a: $y_{mijkl} = \alpha_m + \gamma_l + \eta_{j(k)} + \beta_k + \varepsilon_{i(j(k(m)))}$

where *m* indexes the patient, $\alpha_m \sim N\left(0,\sigma_\alpha^2\right)$ is the random effect for the patient and the error term indicates the variability within observations taken within a treatment period within a block within a patient. The time trend model

Model 3a: $y_t = \alpha_i + \beta t + \gamma X_t + \varepsilon_t$

changes only by having a random intercept $\alpha_i \sim N\left(0,\sigma_\alpha^2\right)$ for patient. These models may be easily extended to encompass interactions between patient and other factors that would indicate variation across patients. In particular, patient characteristics may be able to explain some of the between-patient variance σ_α^2.

If we assume all within block measurements are exchangeable, i.e., that all block-specific treatment effect estimates are similar and can be considered replicates of each other, we can combine results across patients quite simply. First, estimate the treatment effect for patient *i* within block *b* as the difference in the outcomes between treatment 1 and treatment 0, $D_{ib} = Y_{ib1} - Y_{ib0}$. The block-specific treatment effect estimates can then be aggregated across blocks to form the individual treatment effect (ITE) estimate $\bar{D}_i = \sum_{b=1}^{B_i} D_{ib} / B_i$. It is possible to extend this approach into a regression estimate under the broader assumption that allows observed differences across blocks, such as a period effect. The observed individual treatment effects (ITE) $\bar{D}_i$ are unbiased for the true ITE d_i, so that $\bar{D}_i \sim N\left(\delta_i, s_i^2\right)$. The within-patient variance s_i^2 is assumed known and allowed to be specific to each patient (as in a meta-analysis treating the patient as a study). This permits capture of variation in design or implementation of the studies, such as the variation in the number of blocks across patients. For instance, one could assume $S_i^2 = \sigma^2 / B_i$ equals the common within-block variance σ^2 scaled by the number of blocks. If the full model 1 is used, then s_i^2 is estimated from the within-block measurements.

The true ITEs are assumed drawn from a random effects distribution, $\delta_i \sim N(\delta_0, \tau^2)$, where δ_0 denotes the overall mean treatment effect for the population, and τ^2 denotes between-patient variance in the individual mean treatment effects. Prior distributions are placed on the parameters δ_0, τ^2, and σ^2 to represent what is known about these parameters prior to the study. The overall mean treatment effect δ_0 and the individual mean treatment effects δ_i's are estimated using the posterior distribution for each parameter.

The posterior distribution of the patient's ITE, δ_i, provides an opportunity to obtain a more informative estimate of the ITE than is available in a single N-of-1 trial because of the opportunity to borrow strength from the population mean δ_0. Recall that the posterior mean is an average of the sample mean and the prior mean. In this situation, the prior mean δ_0 is the external information coming from other patients and $\bar{D}_i$ is the information coming from the patient. If the patient is like the others, her posterior will be located close to the average

The relationship between individual treatment effect, δ_i, and overall treatment effect, δ_0, depends on the balance between the between-patient variance, τ^2, and the within-patient variance, s_i^2 [20]. When between-patient variance is small compared to within-patient variance (i.e., little or no heterogeneity of treatment effects), the patient-specific mean treatment effects, δ_i, are very similar and close to the posterior mean effect, δ_0. Alternatively, if between-patient variance is large compared to within-patient variance (i.e. strong heterogeneity of treatment effects), the d_i would be estimated to be close to the patient-specific treatment effect estimate, $\bar{D}_i$, with little or no "borrowing from strength." In a sense, the "strength" (population information) to be borrowed does not provide strong statistical information, therefore within-patient information dominates between-patient information.

The model for multiple patients may be extended by considering the model as comprising two parts, within-patient and between-patient. The models for the single N-of-1 trial describe the within-patient parts. The between-patient parts describe factors that vary among patients as in any statistical model with patient units. These include patient characteristics such as co-morbidity, demographics and socioeconomic status. They may also include study and healthcare structure such as the nesting of patients nested within providers and providers within organizations. Each level in the nested structure is represented by a random effect, in addition to the patient level random effect δ_i. For example, the model that accommodates a nested structure with patients nested within practices will have a random

Source: http://www.effectivehealthcare.ahrq.gov/index.cfm
Date Posted: May 20, 2013

effect for practices in addition to a random effect for patients: $\delta_{pi} \sim N\left(\theta_p, \tau_p^2\right)$ with $\theta_p \sim N\left(\theta_0, \omega^2\right)$ where δ_{pi} denotes the individual mean treatment effect for the i^{th} patient in the *p-th* practice, θ_p denotes the mean treatment effect among patients in the p^{th} practice, τ_p^2 denotes the within-practice variance among patients in the p^{th} practice, θ_0 denotes the overall mean treatment effect across practices, and ω^2 denotes the variance across practices. Again, covariates at the practice level can also be incorporated into the model to evaluate the HTE associated with these covariates.

In addition to better estimates of a patient's ITE from borrowing strength from other studies, one also obtains an estimate of the overall treatment effect across patients either as single mean or as a regression. These population effects can be used to inform treatment decisions for similar patients who did not participate in N-of-1 trials.

Finally, when N-of-1 trials with different treatment comparisons are combined across patients, it is possible to consider a network meta-analysis of the N-of-1 trials. Models for network meta-analysis[21,22] incorporate all the pairwise comparisons into a single model for simultaneous estimation. Under assumptions of consistency[23] and similarity[21,24] , direct comparisons of treatments A and B, treatments A and C and treatments B and C may be combined so as to incorporate both their direct estimates and indirect estimates (AC is estimated indirectly through the sum of AB and BC). Such models make optimal use of all the treatment data, leading to more precision in effect estimates as well as the ability to rank treatments. These models hold even when studies do not compare all treatments, but only a subset. For example, a study comparing A and B may be combined with one comparing B and C to get an indirect estimate of A and C. Studies with more than two arms also fit into the model structure. In fact, they provide additional information because their direct and indirect estimates obtained from the same study must be consistent.

5.4 Conclusion

N-of-1 data offer rich possibilities for statistical analysis of individual treatment effects. The more data that are available both within and across patients, the more flexibility models have. This richness does come at the price of the need for careful model exploration and checking. Many errors can be avoided with good study design that respects standard experimental principles and minimizes the risk of complexity caused by autocorrelation as by including washout periods to minimize carryover. Such design and modeling expertise is probably not within the realm of the average clinician and patient undertaking an N-of-1 study. Thus, it is crucial that standard protocols and analyses be available, especially in an automated and computerized format that promotes ease of use and robust designs and models.

Checklist

Guidance	Key Considerations	Check
Treatment assignment needs to be balanced across treatment conditions, using either randomization or counterbalancing, along with blocking	• Design needs to eliminate or mitigate potential confounding effects such as a time trend • Pros and cons of randomization versus counterbalancing need to be considered carefully and selected appropriately. Counterbalancing is more effective if there is good information on critical confounding effect, e.g., linear time trend. Randomization is more robust against unknown sources of confounding. • Blocking helps mitigate potential confounding with time trend, especially when early termination occurs.	□
Blinding of treatment assignment	• Blinding of patients and clinicians, to the extent feasible, is particularly important for N-of-1 trials, especially with self-reported outcomes, when it is deemed necessary to eliminate or mitigate non-specific effects ancillary to treatment • Some non-specific effects might continue beyond the end of trial within the individual patient, therefore should be considered part of the treatment effect instead of a source of confounding	□
Invoke appropriate measures to deal with potential bias due to carryover and slow onset effects	• A washout period is commonly used to mitigate carryover effect • Adverse interaction among treatments being compared indicates the need for a washout period • Absence of active treatment during a washout period might pose an ethical dilemma and diminish user acceptance for active control trials • Washout does not deal with slow onset of new treatment and might actually extend the duration of transition between treatment conditions	□

Source: http://www.effectivehealthcare.ahrq.gov/index.cfm
Date Posted: May 20, 2013

	• Analytic methods can be useful for dealing with carryover and slow onset effects when repeated assessments are available within treatment periods	
Replication of assessments within treatment periods	• Repeated assessments within treatment periods can enhance statistical information (precision of estimated treatment effect) and facilitate statistical approaches to address carryover and slow onset effects • The costs and respondent burden need to be taken into consideration in decisions regarding frequency of assessments	□
Use of adaptive trial design and sequential stopping rule	• Adaptive trial designs and sequential stopping rules can help improve trial efficiency and reduce patients' exposure to the inferior treatment condition	□
Use of appropriate statistical method to analyze outcome data, taking into consideration important features of time series data, including autocorrelation, time trend, and repeated measures within treatment periods	• Mixed effect models, autoregressive models, and dynamic models can be used to analyze time series data from N-of-1 trials • Nonparametric tests are easy to use but might not fully capture time series features • Significance testing is less pertinent for N-of-1 trials than the provision of the information needed for the users to make decisions for future treatments	□
Use of appropriate methods to handle multiple outcomes	• Separate analyses and reporting of trial findings for multiple outcomes can accommodate the patient-centered nature of N-of-1 trials • Explicit pre-specification of weights across outcomes is preferable to post hoc weighting • A composite index or scale can effectively synthesize information across related outcomes and reduce the burden on users to digest trial results across multiple outcomes	□
Presentation of results of statistical analysis in an	• Need to customize format of presentation to needs and preferences	□

informative and user-friendly manner	for individual users • Graphical presentation of trial results is easy to comprehend but might be complicated by autocorrelation, time trend, etc. • Posterior probabilities or odds based on a Bayesian framework is more interpretable for users than p-values based on a frequentist framework	
Borrow from strength	• Bayesian methods can be used to combine data across individuals participating in similar N-of-1 trials, to provide more precise estimates for individual treatment effects, and also to provide estimates for average treatment effects in the population to inform treatment decisions for patients not in the trials • Network meta-analysis can be used to incorporate information from patients whose trials are related to but not identical in design to the treatment conditions compared	□

References

1. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: RandMcNally; 1963.

2. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Belmont, CA: Wadsworth; 2002.

3. Berry SM, Carlin BP, Lee JJ, Muller P. Bayesian adaptive methods for clinical trials. NY: CRC Press; 2010.

4. Zelen M. Play the winner rule and the controlled clinical trial. Journal of the American Statistical Association. 1969;64(325):131-146.

5. Wei LJ, Durham S. The randomized play-the-winner rule in medical trials. Journal of the American Statistical Association. 1978;73(364):840-843.

6. Zucker DR, Ruthazer R, Schmid CH. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. Journal of Clinical Epidemiology. Dec 2010;63(12):1312-1323.

7. Gabler NB, Duan N, Vohra S, Kravitz RL. N-of-1 trials in the medical literature: a systematic review. Medical Care. Aug 2011;49(8):761-768.

8. Schmid CH. Marginal and dynamic regression models for longitudinal data. Statistics in Medicine. 2001;20(21):3295-3311.

9. Hogben L, Sim M. The self-controlled and self-recorded clinical trial for low-grade morbidity. British Journal of Preventive and Social Medicine. 1953 7(4):163-179 Reprinted in International Journal of Epidemiology 2011; 40(6):1438-1454.

10. Zucker DR, Ruthazer R, Schmid CH, Feuer JM, Fischer PA, Kieval RI, Mogavero N, Rapoport RJ, Selker HP, Stotsky SA, Winston E, Goldenberg DL. Lessons learned combining N-of-1 trials to assess fibromyalgia therapies. Journal of Rheumatology. 2006;33(10):2069-2077.

11. Senn S. *Cross-over trials in clinical research.* 2nd ed. Hoboken, NJ: Wiley; 2002.

12. Guyatt GH, Keller JL, Jaeschke R, Rosenbloom D, Adachi JD, Newhouse MT. The n-of-1 randomized controlled trial: clinical usefulness. Our 3-year experience. Annals of Internal Medicine. 1990;112(4):293-299.

13. Larson EB, Ellsworth AJ, Oas J. Randomized clinical-trials in single patients during a 2-year period. Jama-Journal of the American Medical Association. 1993;270(22):2708-2712.

14. McCullough P, Nelder J. *Generalized linear models.* 2nd ed. NY: Chapman and Hall; 1989.

15. Agresti A. *Categorical data analysis.* 2nd ed. Hoboken, NJ: Wiley; 2002.

16. Gill CJ, Savin L, Schmid CH. Why clinicians are natural Bayesians. Bristish Medical Journal. 2005;330(7499):1080-1083.

17. Chaloner K, Church T, Louis TA, Matts JP. Graphical elicitation of a prior distribution for a clinical-trial. Journal of the Royal Stasticial Society. Series D (The Statistician). 1993;42(4):341-353.

18. Zucker DR, Schmid CH, McIntosh MW, D'Agostino RB, Selker HP, Lau J. Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. Journal of Clinical Epidemiology. 1997;50(4):401-410.

19. Duan N, Kravitz R, Schmid C. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. Journal of Clinical Epidemiology. in press.

20. Schmid CH, Brown EN. Bayesian hierarchical models. Numerical Computer Methods, Part C. 2000;321:305-330.

21. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. Research Synthesis Methods. 2012;3(2):80-97.

22. Higgins J, Jackson D, Barrett J, Lu G, Ades A, White I. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. Research Synthesis Methods 2012;3(2):98-110.

23. Lu GB, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. Journal of the American Statistical Association. 2006;101(474):447-459.

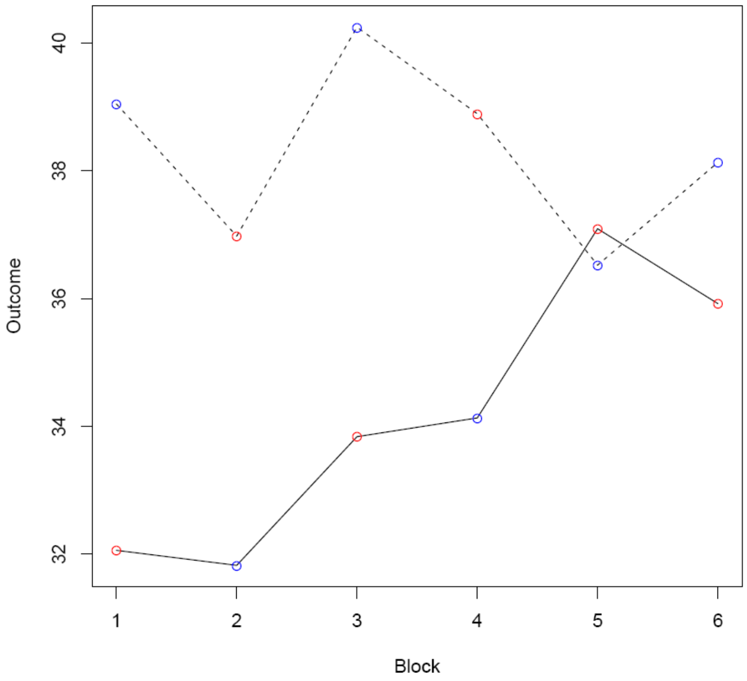
24. Jansen JP, Schmid CH, Salanti G. Directed acyclic graphs can help understand bias in indirect and mixed treatment comparisons. Journal of Clinical Epidemiology. 2012;65(7):798-807.

Table 1. Probability that given outcome or two outcomes together have a treatment effect greater than a given amount

	Outcome		
	1	2	1 and 2
Probability > 0	0.97	0.69	0.67
Probability > 5	0.86	0.50	0.43
Probability > 10	0.51	0.31	0.17
Probability > 15	0.17	0.16	0.02
Probability > 20	0.02	0.07	0.00

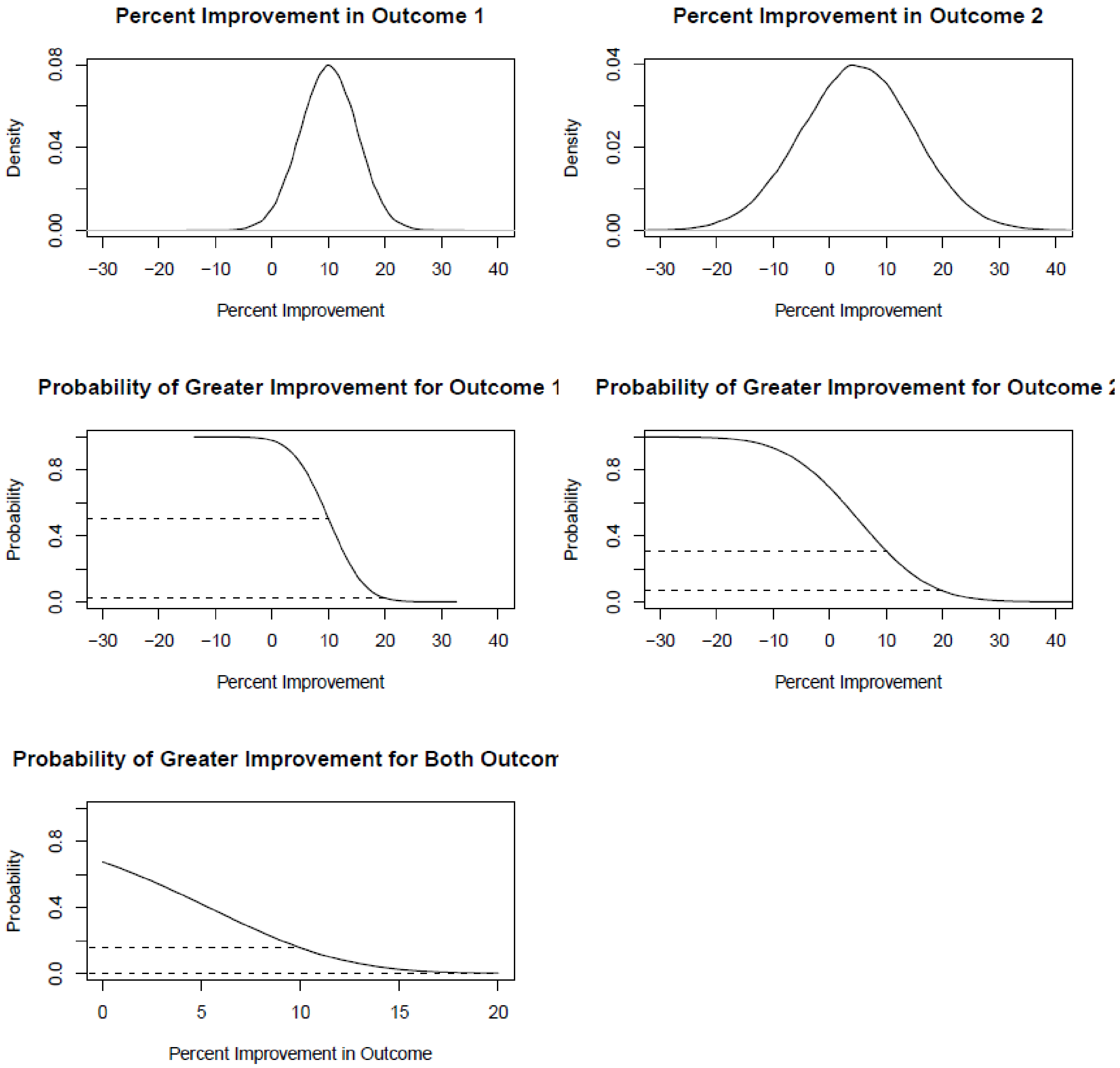
Figures

Figure 1. Data from simulated N-of-1 trial. Two line plots (solid and dotted) show outcomes for two treatments measured within each of 6 blocks. Patients receive each treatment in each block with the point labeled in red taken first.



Source: http://www.effectivehealthcare.ahrq.gov/index.cfm
Date Posted: May 20, 2013

Figure 2. Top row: Posterior distributions in percent improvement (treatment effect) for two outcomes; Middle row: Probability that outcome improves by at least amount on horizontal axis for each outcome; Bottom row: Probability that both outcomes improve by at least amount on horizontal axis.



Percent Improvement in Outcome 1

Percent Improvement in Outcome 2

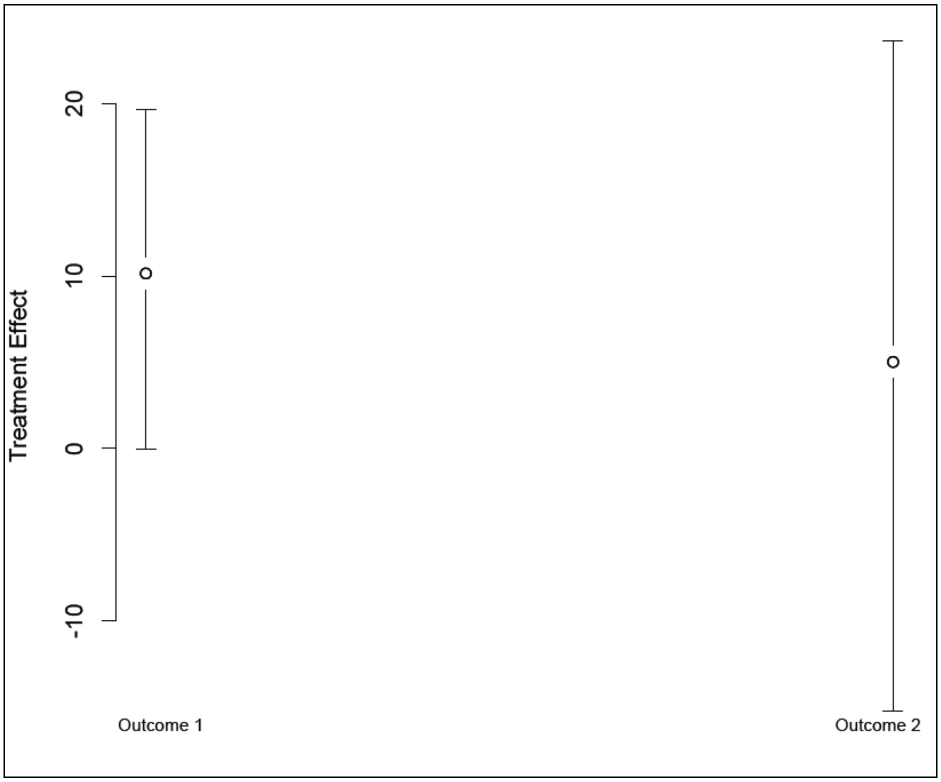
Probability of Greater Improvement for Outcome 1

Probability of Greater Improvement for Outcome 2

Probability of Greater Improvement for Both Outcon

Source: http://www.effectivehealthcare.ahrq.gov/index.cfm
Date Posted: May 20, 2013

Figure 3. Posterior median and 95% central posterior density interval for two outcomes.


Treatment Effect
Outcome 1
Outcome 2